Published online 28 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D797–D802***
doi:10.1093/nar/gkr1127

hiPathDB: a human-integrated pathway database with facile visualization

Namhee Yu[1], Jihae Seo[1,2], Kyoohyoung Rho[1], Yeongjun Jang[1], Jinah Park[1], Wan Kyu Kim[2] and Sanghyuk Lee[1,2,*]

[1]Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 305-806 and [2]Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 120-750, Korea

Received August 15, 2011; Revised October 10, 2011; Accepted November 9, 2011

ABSTRACT

One of the biggest challenges in the study of biological regulatory networks is the systematic organization and integration of complex interactions taking place within various biological pathways. Currently, the information of the biological pathways is dispersed in multiple databases in various formats. hiPathDB is an integrated pathway database that combines the curated human pathway data of NCI-Nature PID, Reactome, BioCarta and KEGG. In total, it includes 1661 pathways consisting of 8976 distinct physical entities. hiPathDB provides two different types of integration. The pathway-level integration, conceptually a simple collection of individual pathways, was achieved by devising an elaborate model that takes distinct features of four databases into account and subsequently reformatting all pathways in accordance with our model. The entity-level integration creates a single unified pathway that encompasses all pathways by merging common components. Even though the detailed molecular-level information such as complex formation or post-translational modifications tends to be lost, such integration makes it possible to investigate signaling network over the entire pathways and allows identification of pathway cross-talks. Another strong merit of hiPathDB is the built-in pathway visualization module that supports explorative studies of complex networks in an interactive fashion. The layout algorithm is optimized for virtually automatic visualization of the pathways. hiPathDB is available at http://hiPathDB.kobic.re.kr.

INTRODUCTION

Pathways are the essential units of biological processes and embody the regulatory steps of cellular activities. According to Pathguide, a resource that compiles the resources related to pathways and molecular interactions, there are 59 pathway-related resources and 151 166 pathway entries within (1). KEGG, Reactome, BioCyc are among the best-known primary databases developed and maintained by a few dedicated research groups (2–4). BioCarta (http://www.biocarta.com) and WikiPathways are the examples of community-based resources depending on experts in specific subjects (5,6).

With multiple pathway databases in distinct formats available, data integration has become an important issue in utilizing these resources systematically and efficiently. Pathway interaction database (PID) from the NCI-Nature resources and Pathway Commons from the MSKCC and the University of Toronto groups are two notable examples of pathway integration (7,8). Importantly, integrations of pathways extend the knowledge significantly beyond that contained within individual pathway. For example, the TNF signaling pathway in Reactome contains only 5 interactions, whereas NCI-Nature PID integrating BioCarta with Reactome contains 28 interactions.

The critical issue in pathway data integration is establishing a standard format for exchange. Even though the Biological Pathways Exchange (BioPAX) and Systems Biology Markup Language (SBML) have been proposed as standards for pathway data exchange in the XML format (9,10), pathway integration is still a non-trivial task since each pathway database was constructed based on its own pathway model. In other words, the data models reflect the researcher's view on pathways and are often incompatible with simple integration by compilation and format unification. Because of these difficulties, most integrated pathway databases are simple collections of

*To whom correspondence should be addressed. Tel: +82 42 879 8500, +82 42 879 8511; Fax: +82 42 879 8519; Email: sanghyuk@kribb.re.kr

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

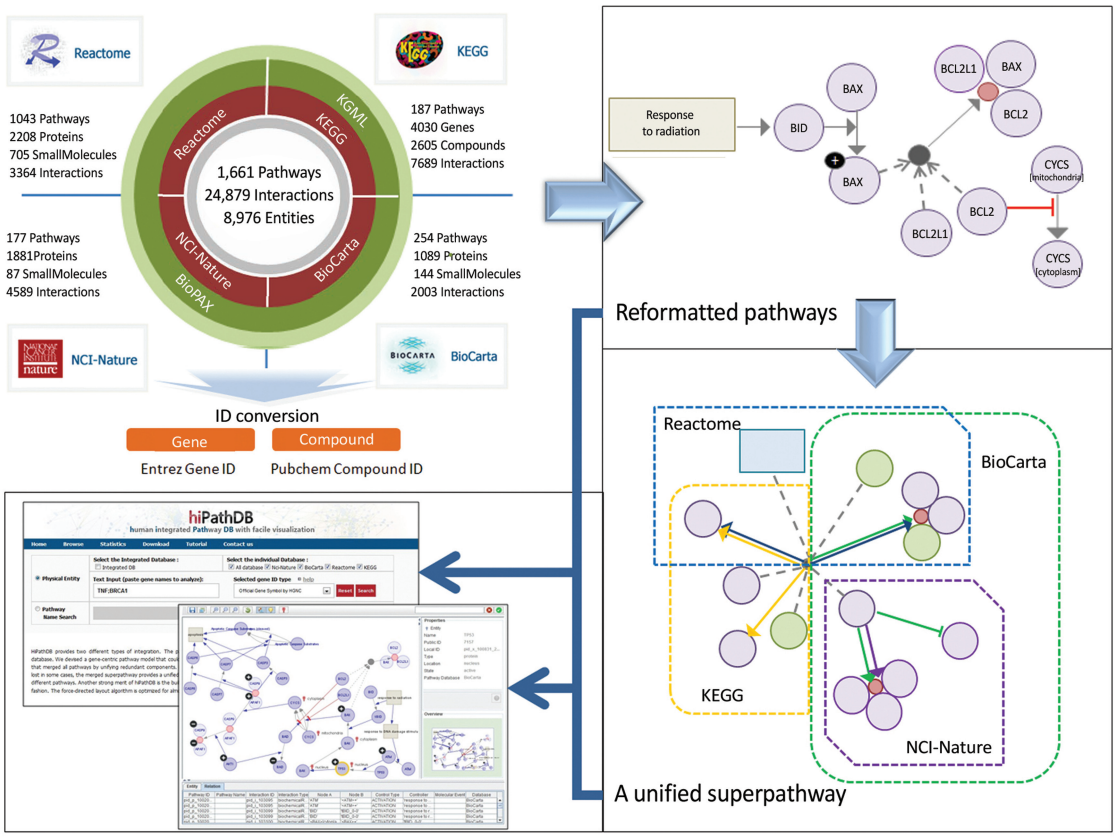

Figure 1. System overview of hiPathDB.

existing databases in the original format without serious remodeling. Such limitation notwithstanding, they still enable identification of all pathways in different original databases with a single gene or keyword search. For example, both PathJam and HPD cover four pathway databases (NCI-Nature PID, Reactome, BioCarta and KEGG), but the integration is done at the level of gene lists, not interactions (11,12). This makes it difficult to analyze or visualize pathways from different databases with a coherent and comprehensive perspective.

In principle, pathways can be defined as the compiled relationships among the biological entities in various given contexts. Although biologists typically focus on limited portions related to specific biological phenomena, it is often the case that pathways previously thought independent turn out to be interconnected. Consequently, as more regulatory relationships accumulate, pathway integration at the molecular level becomes increasingly informative. It is thus necessary to develop means to explore the pathway space efficiently. Commercial software such as Ingenuity Pathway Analysis or MetaCore collects such information from the literature by manual curation but has disadvantage of inherent inefficiency as well as high costs for the access.

Visualization of pathways is another challenge in the pathway analysis. Pathway resources usually provide diagrams drawn manually with no allowance of interactive editing (e.g. KEGG and BioCarta). While they provide informative knowledge compiled by experts, an interactive interface for browsing and extending part of pathways is missing. Most programs for pathway visualization are stand-alone applications mainly for editing purposes, as can be seen in the example of GenMAPP and PathVisio (13,14). Combining integrated pathway data with a powerful interactive visualization software is necessary to help biologists obtain insights for underlying molecular mechanisms. Currently, ConsensusPathDB (CPDB) is the only pathway database in this category, but their graphical interface is too complicated for facile interpretation (15).

Here, we introduce a new integrated pathway database with effective visualization support for interactive pathway modeling. Integration at both pathway and entity levels are achieved for four of the most widely used pathway databases: NCI-Nature PID, Reactome, BioCarta and KEGG. The key advantage is the utilization of an elaborate pathway model which accommodates complex biological functions with minimum loss of detailed information.

SYSTEM OVERVIEW

The overview of hiPathDB workflow is shown in Figure 1. Pathway data were collected from four different sources: NCI-Nature PID, Reactome, BioCarta and KEGG.

In total, 1661 pathways consisting of 24 879 interactions among 8976 distinct physical entities are included.

Each signaling step from each of the databases was re-modeled in a coherent fashion rather than simply compiled. Detailed description of our pathway modeling is provided in the following section. Entities in the reformatted pathways were integrated at the molecular level to create a unified superpathway. This allows users to explore the entire network of molecular pathways. Detailed information on data integration procedure is provided in the website online documentation.

The website supports both gene and pathway queries. A Java applet for pathway visualization was developed for interactive visualization and for editing of the reformatted pathways and unified superpathway. A stand-alone in-house program MONGKIE, dedicated for pathway visualization, was modified to be implemented as an applet.

DATA IMPORT AND PATHWAY REMODELING

As most pathway databases support data exchange in the BioPAX format, we adopted BioPAX (level 2) as the standard format for data collection. However, it should be noted that a simple binary model of nodes and edges cannot describe the cellular processes properly. For example, physical entities in the BioPAX data model can be any of DNA, RNA, protein and complex. In addition, molecular interactions cover diverse types of processes such as biochemical reaction, complex assembly, transport, catalysis and modulation. Therefore, we needed to devise an elaborate pathway and database model to integrate pathways from different sources with minimal loss of information. The database schema of entity-relation diagram reflecting these characteristics is provided in the website documentation.

Data for NCI-Nature curation and BioCarta were downloaded from the NCI-Nature PID database. The Reactome (release version 32) was obtained from the website. The downloaded XML files were parsed by using our in-house Python parsers for BioPAX data. Parsed data were converted according to our own pathway data model and stored into the relational database. Manual curation was necessary for nodes without cross-reference information. The KEGG database also provides their data in the BioPAX format, but the information content in this format was only partial. Thus, we downloaded the data in the KGML format and converted them into the BioPAX format, which required careful comparison of the two models to identify controller nodes. Several additional data such as enzyme information were also obtained using the KEGG API.

For visualization, we introduced the concept of 'abstract node' to be compliant with the BioPAX entity concept as shown in Figure 2. The abstract node can be a usual node, a pseudo node, or even a complex node. As exemplified in Figure 2, the pseudo node (gray circle) and complex node (red circle) can be used to describe a physical or functional aggregation of BAX, BCL2L1 and BCL2. Our data model also supports activation (BAX) and translocation (CYCS) processes.

All pathways in the four pathway databases were imported and reformatted according to the procedure as described above. Figure 3 shows the pathway visualization of the reformatted pathway in hiPathDB for the BioCarta 'apoptotic signaling in response to DNA damage' pathway. Visualization was done using the website applet. Comparison of this image with the original BioCarta image available at http://www.biocarta.com/pathfiles/h_chemicalPathway.asp readily reveals that our modeling captures most of the important biological details in the original database: the biological process indicated as 'response to radiation' (upper right) promotes BID cleavage which activates BAX; DNA damage causes activation of ATM-TP53 signaling which is followed by BAX transcription and translocation into the cytoplasm; activated BAX promotes cytochrome C (CYCS) release from mitochondria to cytoplasm by forming a complex with BCL2 and BCL2L1 which in turn inhibits the release of CYCS. Further, it can be seen that CYCS subsequently activates a complex consisted of CASP9 and APAF-1 (a protease released from mitochondria), which triggers the caspase cascade by cleaving and activating caspases (CASP3, CASP6, CASP7) ultimately leading to apoptosis (upper left). That AKT1 activation inhibits the whole apoptosis process by inactivating both CASP9-APAF-1 complex and BAD protein is also incorporated as a part of the overall process.

This example demonstrates that the full details of apoptosis process can be gleaned from our reformatted pathway in an intuitive manner. It is noteworthy that our pathway diagram, produced almost automatically with minor interactive modifications, is as informative as the original BioCarta picture that was drawn manually by experts in the field.

UNIFIED SUPERPATHWAY

We also implemented pathway integration at the entity level. In this gene-centric model, the entities are mapped to the standard identifier, merged and unified as non-redundant data set. These non-redundant set of nodes are interconnected into a single unified pathway like the superpathway used in BioCyc. The resulting unified pathway can be used to explore the molecular network and to identify pathway cross-talks efficiently.

As genes, proteins and chemicals often have multiple synonyms and homographs where two or more different biochemical entities have the same symbol, unique standard identifiers for the entities are required. We adopted the Entrez GeneID for genes and their products, PubChem compound IDs for chemicals. Mapping was processed using cross-references, and the process labels were used as the identifiers for the relevant biological processes.

In the entity-level integration, it is virtually impossible to retain the biological details of pathways. Therefore, we converted the pathway model in the BioPAX format into

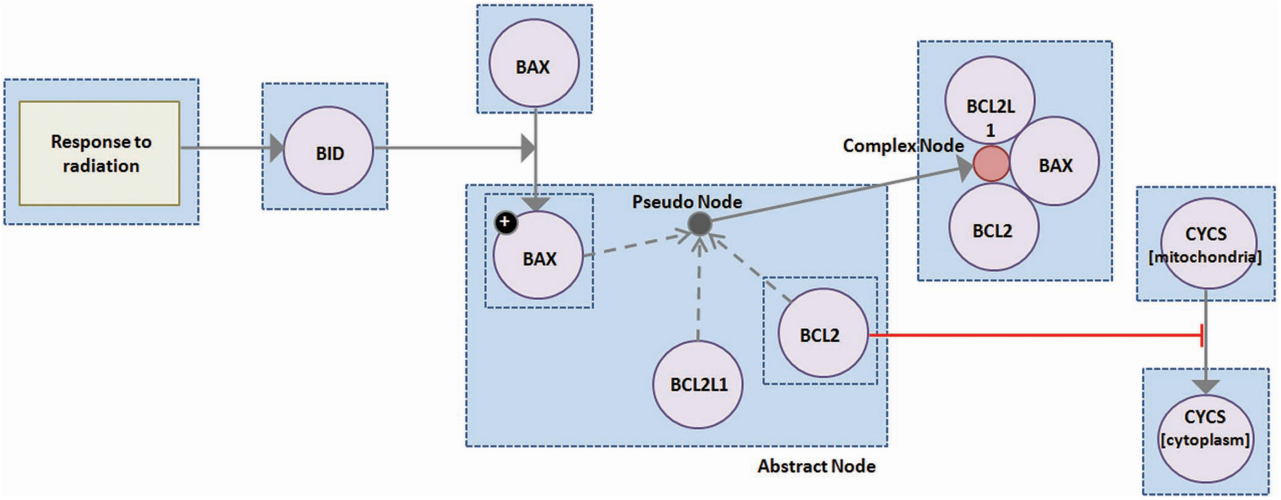

Figure 2. Data model for pathway integration. See the text for description of the abstract node, pseudo node (gray circle) and complex node (red circle).

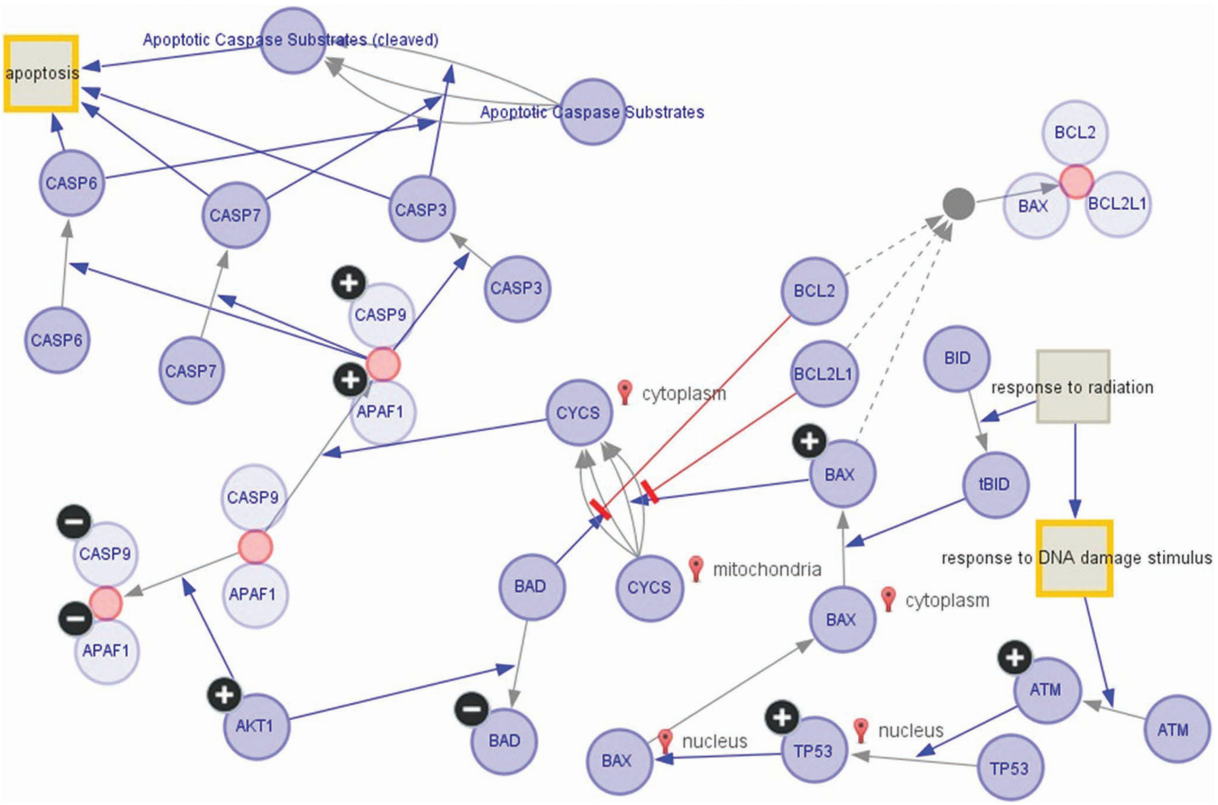

Figure 3. Reformatted modeling of the BioCarta 'apoptotic signaling in response to DNA damage' pathway. Note that + and − next to the gene circles indicate the active and inactive forms, respectively. Small red ovals indicate translocation, and the new location of protein is provided next to the oval (yellow box, biological processes; blue circle, gene or protein; gray circle, pseudo node; red circle, complex node).

the Simplified Interaction Format (SIF) as described in the Pathway Commons resource with minor modifications of the rules. This conversion led to representation of all pathway interactions as binary pair-wise interactions, suitable for network analysis, visualization and navigation. Because the molecular details are already lost, we chose not to discriminate the gene products with different active states or cellular locations to increase the number of equivalent entities.

As an illustration for the utility of the unified superpathway, we again examined the BAX-mediated apoptosis process. Figure 4 shows a part of the unified pathway related to BAX, BCL2 and CYCS. Binary networks centered around BAX (lower right) were

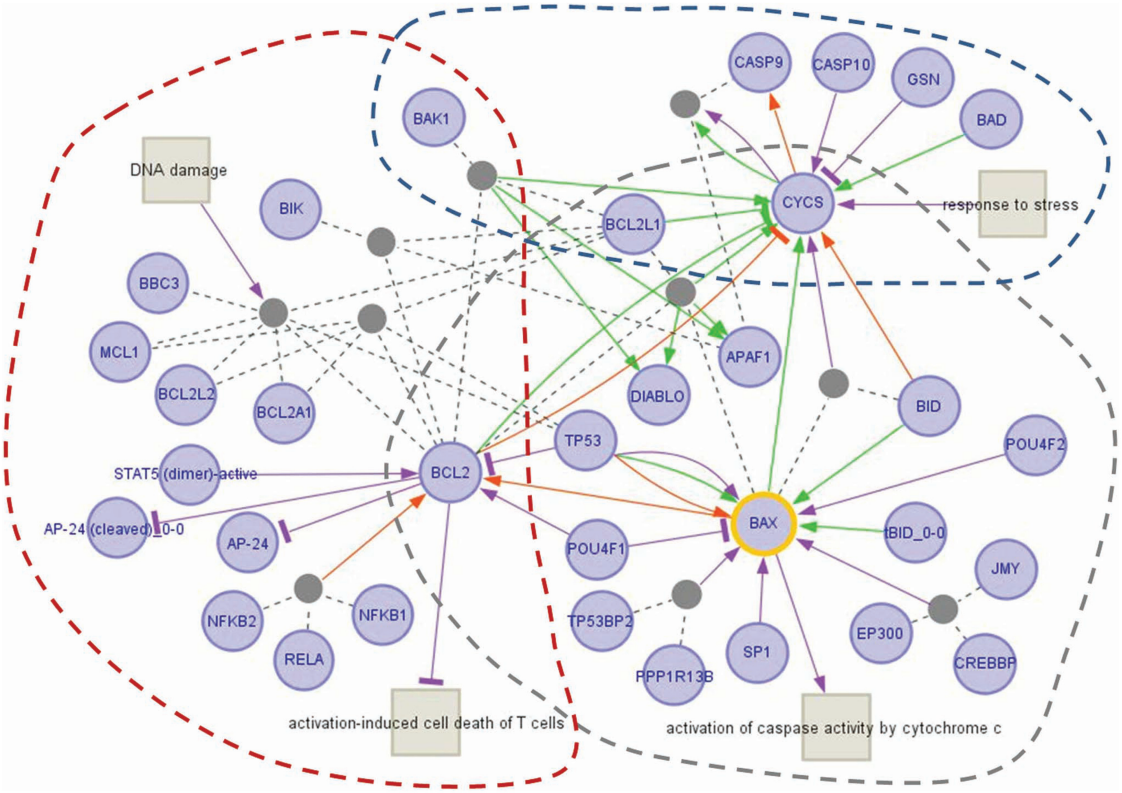

Figure 4. Exploration of unified superpathway to identify pathway cross-talks. Edge colors indicate the source database (green, BioCarta, orange, KEGG, violet, NCI-Nature PID). The gray-dotted area (lower right) shows the query result on BAX. Blue-dotted area (upper right) is the node expansion at CYCS, and the red-dotted area is the node expansion at BCL2 gene.

expanded at nodes BCL2 and CYCS. The key features of apoptosis process due to DNA damage remain essentially identical to the process illustrated in the BioCarta pathway (Figure 3; 'DNA damage'-TP53-BAX-BCL2-CYCS-caspases). However, the unified pathway shows several additional nodes from the NCI-Nature PID and Reactome databases. Among them, NFkB, CREBBP and SP1 represent the new nodes with important regulatory functions in various cellular processes. The NCI-Nature PID indicates that the stress response also leads to apoptosis via CYCS. Expanding at other important nodes would display numerous other genes and processes related to BAX, BCL2 and CYCS. Even though the details of molecular mechanism are lost in the unification process, it is evident that the unified superpathway is useful to survey the molecular network in more comprehensive way and to identify key regulators and pathway cross-talks.

WEB INTERFACE AND IMPLEMENTATION

The hiPathDB website incorporates various user-friendly features. Searches for physical entities can be carried out either for unified superpathway or for the remodeled pathways. Diverse ID types can be entered in the query window. Querying multiple gene names gives all the pathways, entities and edge relations organized in an intuitive way. Pathway search supports the auto-complete mode in the input process to enable users to find and select queries quickly from the recommended list. Most useful features available in other pathway databases are implemented as well to enhance user convenience.

The visualization tightly integrated with the database is the strongest merit of hiPathDB. The software supports diverse types of node–edge relations such as aggregation, complex formation, translocation and biochemical reactions. The force-directed layout algorithm is optimized for pathway visualization. In addition to the basic operations of zooming and panning, coherently working multiple windows (pathway, overview, description and information windows) facilitate network navigation and information extraction substantially. Users can delete nodes directly within network window to reduce network complexities. hiPathDB also supports node expansion in the unified superpathways.

The response time is an important issue in supporting interactive visualization. To reduce the response time, we made two additional mirror sites in US/Canada and Europe regions using the Amazon Web Services (http://us.hipathdb.kobic.re.kr and http://eu.hipathdb.kobic.re.kr). The amount of data transfer was minimized as well.

Even though the current visualization applet is one of the most powerful programs, the MONGKIE program has several other modules useful for pathway analysis. This includes the network clustering, expression overlay, etc. For users inclined to utilize features in the

MONGKIE, hiPathDB support exporting pathway diagrams in the GraphML format.

CONCLUSION

hiPathDB is an integrated database for human pathways with a state-of-the-art visualization. Our data modeling maintains the molecular details of signaling processes. The unified superpathway achieved by entity level integration facilitates network analysis and navigation. With these unique features, hiPathDB should be a valuable addition for communities interested in analyzing pathways. There are many aspects to improve, and we plan to add other pathway and interaction databases in the near future. The website and visualization software will be improved to enhance user convenience based on community feedbacks.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the dedicated effort of curation teams that are creating and sharing biological pathway information. They also appreciate Ms Sun Young Ji and Mr Ji Han Kim for helping the website construction.

FUNDING

KRIBB Research Initiative Program; National Research Foundation of Korea (NRF) grants funded by the Korea government (MEST) (No. 2011-0002321, No. 2011-0019745, No. R15-2006-020); GIST Systems Biology Infrastructure Establishment Grant (2011) through Ewha Research Center for Systems Biology (ERCSB). Funding for open access charge: KRIBB Research Initiative Program.

Conflict of interest statement. None declared.

REFERENCES

1. Bader,G.D., Cary,M.P. and Sander,C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.
2. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
3. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
4. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
5. Nishimura,D. (2001) BioCarta. *Biotech Softw. Internet Rep.*, **2**, 117–120.
6. Pico,A.R., Kelder,T., Iersel,M.P., van Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
7. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.
8. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
9. Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotech.*, **28**, 935–942.
10. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
11. Glez-Peña,D., Reboiro-Jato,M., Domínguez,R., Gómez-López,G., Pisano,D.G. and Fdez-Riverola,F. (2010) PathJam: a new service for integrating biological pathway information. *J. Integr. Bioinformatics*, **7**, 1–17.
12. Chowbina,S.R., Wu,X., Zhang,F., Li,P.M., Pandey,R., Kasamsetty,H.N. and Chen,J.Y. (2009) HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics*, **10(Suppl. 1)**, S5.
13. Salomonis,N., Hanspers,K., Zambon,A.C., Vranizan,K., Lawlor,S.C., Dahlquist,K.D., Doniger,S.W., Stuart,J., Conklin,B.R. and Pico,A.R. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.
14. Iersel,M.P., van Kelder,T., Pico,A.R., Hanspers,K., Coort,S., Conklin,B.R. and Evelo,C. (2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics.*, **9**, 399.
15. Kamburov,A., Pentchev,K., Galicka,H., Wierling,C., Lehrach,H. and Herwig,R. (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.